2011

Measurement of patient-reported outcomes in surgical orthopaedics

Kimberly Joy Neufeld

Measurement of patient-reported outcomes in surgical orthopaedics

(Spine title: Measurement of patient-reported outcomes)

(Thesis format: Monograph)

by

Kimberly Joy Neufeld

Graduate Program in Health and Rehabilitation Sciences
Measurement and Methods Field

A thesis submitted in partial fulfilment
of the requirements for the degree of
Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

THE UNIVERSITY OF WESTERN ONTARIO
THE SCHOOL OF GRADUATE AND POSTDOCTORAL STUDIES

CERTIFICATE OF EXAMINATION

Supervisors	Examiners
Dr. Dianne Bryant	Dr. Trevor Birmingham
Dr. Andrew Johnson	Dr. Chris Lee
	Dr. Chris Hyson

The thesis by

Kimberly Joy Neufeld

entitled:

Measurement of patient-reported outcomes in surgical orthopaedics

is accepted in partial fulfilment of the
requirements for the degree of
Master of Science

Date_____

_____
Chair of the Thesis Examination Board

Abstract

We examined the predictive relationship between factors previously shown to relate to patients' ability to accurately recall past health states and discrepancies between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status.

We used the least squares method of multiple linear regression to construct models that used the absolute and simple differences between pre-surgical and post-surgical ratings of physical health, mental health, function, and quality of life as dependent variables.

Four hundred and eleven patients were included in the analyses. Individually and in combination, the selected factors explained a limited amount of variance in the absolute and simple differences between ratings.

Age, sex, mental state at the time of recall, health status at the time of recall, and time between ratings are not strong predictors of discrepancies between pre-surgical and post-surgical ratings of pre-intervention health status in the early post-surgical period.

Keywords: Clinical trials, interventions, surgery, orthopaedics, data collection, baseline, prospective, retrospective, recall

Co-Authorship Statement

The idea for conducting research on this methodological issue came from Dianne Bryant's experiences working as a coordinator on clinical trials within the field of surgical orthopaedics. I obtained data sets from Dianne Bryant and her other students (Jacquelyn Marsh and Jennifer Gow). I combined these data sets, focused the research question, and designed the original plan for analysis. This plan was progressively refined through discussions with Dianne Bryant and Andrew Johnson. I was solely responsible for conducting the analyses. I wrote the original draft of this thesis (including interpretation of the statistical results) and sent it to Dianne Bryant and Andrew Johnson for comments and suggestions.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Measurement of self-reported health status is one of the preferred methods for assessing patient outcomes. Many tools have been developed to measure patients' perceptions of their physical health, mental health, function, and quality of life. When conducting clinical trials to test the effectiveness of health-related interventions, researchers often use these tools to obtain ratings of pre-intervention (baseline) health status. Such data are used to demonstrate similarity between treatment groups and to control for differences in health status that existed prior to intervention.

In clinical trials involving surgical interventions, research staff frequently collect baseline data from patients who – following pre-surgical evaluation – appear to satisfy eligibility criteria, yet have the potential to be disqualified following intra-surgical evaluation. When the diagnostic accuracy of tests used in the pre-surgical evaluation is low, the number of patients who prove to be ineligible following intra-surgical evaluation can be high. This creates huge inefficiencies in the data collection process.

A recent clinical trial comparing the effectiveness of two approaches to repairing meniscal tears (inside-out sutures versus bioabsorbable arrows) in patients undergoing arthroscopy illustrates the inefficiencies introduced by the inability to fully assess patient eligibility prior to surgery (Bryant, Dill, Litchfield, Amendola, Giffin, & Kirkley, 2007). In this trial, only 100 of the 700 patients who consented to participate and provided pre-surgical (prospective) ratings of pre-intervention health status were found to have a lesion

amenable to repair using either approach upon intra-surgical evaluation. Thus, 86% of consented patients who completed pre-surgical assessments were excluded from the study post-surgically. Collection of baseline data on these patients increased patient burden (600 patients completed an unnecessary 40 minute assessment) and wasted approximately $12,000 CAD in research funds (based on 400 hours of research assistant time reimbursed at a rate of $30 CAD per hour).

This problem is not unique to that study. In fact, it is a common phenomenon not only in research on surgical interventions for patients with knee problems, but also in research on surgical interventions for other patient populations. For example, recent reports from ongoing clinical trials comparing the effectiveness of different approaches to repairing rotator cuff tears indicated that 29% (93 out of 317) and 38% (38 out of 100) of participants who provided pre-surgical (prospective) ratings of pre-intervention health status were found to be ineligible following intra-surgical evaluation (Bryant, Litchfield, Holtby, Willits, Drosdowech, Spouge, & Guyatt, n.d.; MacDermid, Holtby, Razmjou, Bryant, & JOINTS Canada, 2006). When one considers the costs incurred in these and similar studies, one can understand why it is important for methodologists to develop strategies to increase the efficiency of data collection in this type of research.

Substitution of post-surgical (retrospective) ratings of pre-intervention health status for ratings that would have otherwise been collected pre-surgically (prospectively) could increase the efficiency of data collection when patient eligibility cannot be fully determined prior to intra-surgical evaluation. This strategy, however, necessitates a high

level of agreement between these two types of baseline data. In other words, patients must be able to accurately recall past health states and must provide similar ratings of pre-intervention health status at different points in time (before and after surgery).

Theories of memory processes (e.g., Ross, 1989; Sprangers & Schwartz, 1999) predict that there are conditions under which discrepancies between ratings of specific health states will occur. Differences in the magnitude and direction of discrepancies between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status in clinical trials may be linked to identifiable factors, such as patient characteristics. They may also be linked to aspects of trial design. If such differences are large, use of post-surgical (retrospective) ratings of pre-intervention health status may be more suitable in some studies than others. Thus, it is important to identify predictors of discrepancies between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status.

Previously, we demonstrated – in three independent randomized controlled trials – that the level of agreement between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status was sufficiently high to recommend substituting pre-surgical ratings with post-surgical ratings in patients undergoing knee, hip, and shoulder surgery (Bryant, Norman, Stratford, Marx, Walter, & Guyatt, 2006; Gow, 2010; Marsh, Bryant, & MacDonald, 2009). Nevertheless, agreement was not perfect. We were unable to identify significant predictors of discrepancies between pre-surgical (prospective) and post-surgical (retrospective) ratings in patients

undergoing knee surgery (Bryant, 2005) and hip surgery (Marsh, Bryant, & MacDonald, n.d.). However, our ability to do so was limited by lack of variability in specific independent variables (such as time between ratings), as well as lack of variability in the dependent variables (absolute and simple differences between ratings). These limitations were a result of standardized protocols that limited variability within the individual trials. In this thesis, we address these limitations by combining data sets from all three trials. These varied sufficiently in their protocols to create additional variability within each of these variables.

Chapter 2: Literature Review

2.1 Chronic health conditions and musculoskeletal problems

Chronic health conditions are physical or mental health problems that persist despite treatment and require ongoing management over an extended period of time (Nasmith, Ballem, Baxter, Bergman, Colin-Thomé, Herbert, Keating, Lessard, Lyons, McMurchy, Ratner, Rosenbaum, Tamblyn, Wagner, & Zimmerman, 2010). Such conditions are the leading cause of death in nearly all countries (World Health Organization [WHO], 2005). In countries with older demographics and more advanced health care (such as Canada), they are also the leading cause of disability and use of health care resources (Nasmith et al., 2010). While some people in these countries have only one chronic health condition, many have comorbidities (problems related to an initial condition) or multimorbidities (multiple problems, some related to each other, some complicating each other, and some unrelated to each other but coexisting) (Canadian Orthopaedic Care Strategy Group [COCSG], 2010; Nasmith et al., 2010). In Canada, approximately 50% of the population has at least one chronic health condition (Advisory Committee on Population Health, 2002).

Chronic musculoskeletal problems, such as arthritis, back pain, and osteoporosis, are the most prevalent type of chronic health conditions in Canada (COCSG, 2010). Such problems are generally accompanied by pain, fatigue, emotional disturbances, and reductions in physical and social function (Murphy, Spence, McIntosh, & Connor

Gorber, 2006). Acute musculoskeletal problems (injuries) also affect a substantial proportion of Canadians (especially those aged 44 years or younger) and are a risk factor for the development of chronic musculoskeletal problems later in life (COCSG, 2010).

2.2 Health care reform and evidence-based decision-making

In recognition of the costs associated with chronic health conditions, the question of how to most effectively prevent and manage such conditions has become increasingly relevant (Nasmith et al., 2010; WHO, 2005). The ultimate goal of health care reform is to establish a system in which the right care is consistently delivered in the right location by the right care provider at the right time (COCSG, 2010; Nasmith et al., 2010). In such a system, decisions regarding what constitutes the best care, location, care provider, and time should be based on sound scientific data (evidence-based).

In Canada, primary care physicians and nurse practitioners often assume a lead role in caring for those with both acute and chronic musculoskeletal problems. Such care providers may recommend general lifestyle modifications (including changes in nutrition and physical activity) or pharmaceutical interventions (including analgesics and anti-inflammatory medications). In some cases, they may also refer patients to allied care providers (such as physical or occupational therapists) or specialist physicians (such as rheumatologists or orthopaedic surgeons).

Orthopaedic surgeons are the type of specialist most frequently seen by patients with musculoskeletal problems (COCGS, 2010). Although they may recommend the same or similar non-surgical means to address musculoskeletal problems as other care providers, their unique contribution to health care relates to their ability to address structural issues using surgical interventions, such as arthroscopy (visual inspection of a joint, with or without repair of damaged cartilage or ligaments) and arthroplasty (joint replacement).

Despite increased demand for clinicians to engage in evidence-based decision-making (Sackett, Rosenberg, Muir Gray, Haynes, & Richardson, 1996), many of the interventions used by orthopaedic surgeons are introduced into practice without rigorous evaluation (Busse & Heetveld, 2006). In part, this reflects the fact that, unlike pharmaceutical interventions, regulatory bodies (such as those of the U.S. Food and Drug Administration or Health Canada) do not mandate rigorous evaluation of surgical interventions. The deficiency of high quality research can also be attributed to ethical, logistical, and methodological challenges associated with conducting clinical trials involving surgical interventions (Busse & Heetveld, 2006; Chung & Burns, 2008).

2.3 Clinical trials

Clinical trials are prospective research studies that evaluate the effects of one or more health-related interventions on human participants (Canadian Institutes of Health Research, n.d.). They may be classified as therapeutic or preventive (depending on whether they examine the effectiveness of interventions in reducing or eliminating

distressing symptoms associated with existing health problems or in preventing new problems from arising) and are often described according to their design characteristics (including the level of control over extraneous variables and the procedures used to assign participants to groups) (Portney & Watkins, 2009).

All clinical trials evaluate the effects of interventions by examining between-group differences in outcomes. Outcomes can be assessed in many different ways, including direct observation or collection of patient self-reports. A common way to collect self-report data is by using patient-reported outcome measures.

2.4 Patient-reported outcome measures

Patient-reported outcome measures are measurement tools used to gain insight into the way that patients perceive their health and the impact that changes in health status have on multiple facets of their lives. In recent years, the use of such tools in orthopaedic research has increased, particularly in randomized controlled trials (Beaton & Schemitsch, 2003). Preference for a multi-faceted approach to the measurement of outcomes can be viewed as part of the growing acceptance among clinical and research communities of a model of health that emphasizes not only the biological aspects of health (the primary focus in the previously dominant model), but also the psychological and social aspects (Borrell-Carrió, Suchman, & Epstein, 2004; Portney & Watkins, 2009; WHO, 2001). Furthermore, many studies have demonstrated discrepancies between clinicians' evaluations and patients' evaluations of health status (Sprangers & Aaronson,

1992). Preference for patient-reported outcomes over clinician-reported outcomes can be viewed as part of the growing recognition that changes in physiologic endpoints (measured in terms of physical examination findings, laboratory values, and the results of functional tests) bear a limited relation with changes in self-reported health status making them inappropriate surrogates for patient-important endpoints (Bryant, Schünemann, Brozek, Jaeschke, & Guyatt, 2007).

According to the Scientific Advisory Committee of the Medical Outcomes Trust (2002), there are eight key attributes of patient-reported outcome measures: the conceptual and measurement model, reliability, validity, responsiveness, interpretability, respondent and administrative burden, alternate forms, and cultural and language adaptations. When conducting a clinical trial, understanding of the attributes of relevant tools should be used – together with knowledge of the conditions and interventions of interest – to guide the selection of the most appropriate measurement tools (Bryant & Fernandes, 2011).

2.5 Discrepancies between ratings of specific health states made at different points in time

As described in Chapter 1, in clinical trials involving surgical interventions, substitution of post-surgical (retrospective) ratings of pre-intervention health status collected for ratings that would have otherwise been collected pre-surgically (prospectively) could increase the efficiency of data collection when patient eligibility cannot be fully determined prior to intra-surgical evaluation. This strategy, however, necessitates a high level of agreement between these two types of baseline data. In other words, patients

must be able to accurately recall past health states and must provide similar ratings of pre-intervention health status at different points in time (before and after surgery).

Theories of memory processes predict that there are conditions under which discrepancies between ratings of specific health states (such as pre-intervention health status) will occur. Key examples of such theories are those described by Ross, Sprangers, and Schwartz.

According to Ross (1989), people possess implicit theories regarding the inherent consistency of their attributes, as well as the conditions that are likely to promote stability or change. When asked to make a judgment about past personal attributes, people assess their current status and use their implicit theories to infer what their status was in the past. Discrepancies between ratings of specific health states made at different points in time are expected to occur when people adopt a theory of stability under conditions of actual change, when they adopt a theory of change under conditions of actual stability, or when they adopt a theory of change that is different from actual change.

According to Sprangers and Schwartz (1999), people infer what their past status was using their current internal standards of measurement, values, and definitions of the attributes. Any or all of these may have changed as a result of new information acquired since making earlier judgments about the same attributes. Discrepancies between ratings of specific health states made at different points in time are expected to occur under

conditions that promote changes in internal standards of measurement, values, and/or definitions of health status constructs.

Previous studies examining patients' ability to accurately recall past health states have produced mixed results. Some authors claim that patients can accurately recall past health states (e.g., Singer, Kowalska, & Thode, 2001), while others claim they cannot (e.g., Mancuso & Charlson, 2005). Furthermore, at the time of recall, some studies have shown patients to rate past health states as being better than reported at an earlier point in time (e.g., Eich, Reeves, Jaeger, & Graff-Radford, 1985), while others have shown the opposite trend (patients rate past health states as being worse than reported at an earlier point in time) (e.g., Everts, Karlson, Währborg, Abdon, Herlitz, & Hedner, 1999).

Differences in the magnitude and direction of discrepancies between ratings of specific health states made at different points in time may be linked to characteristics of patients, interventions, or trial design.

2.6 Characteristics of patients as predictors of discrepancies between ratings

Age, sex, and health status at the time of recall are among the characteristics of patients that have been shown to relate to patients' ability to accurately recall past health states. Several studies have shown the magnitude of the discrepancy between ratings of specific health states made at different points in time to be larger in those who are in a worse health state at the time of recall (e.g., Jamison, Sbrocco, & Parris, 1989) or female (e.g.,

Hunter, Philips, & Rachman, 1979). Several studies have also shown that, at the time of recall, patients who are in a better health state at the time of recall (e.g., Kent, 1985) or are female (e.g., Mancuso & Charlson, 2005) tend to rate past health states as being better than reported at an earlier point in time. Such trends, however, are not consistent across all studies. For example, although studies in adults tend to show the magnitude of the discrepancy between ratings to be larger in those who are older (e.g., Lingard, Wright, Sledge, & the Kinemax Outcomes Group, 2001), studies in children tend to show the opposite trend (e.g., Zonneveld, McGrath, Reid, & Sorbi, 1997).

Another characteristic of patients that has been shown to relate to their ability to accurately recall past health states is the nature of their health condition. Specifically, studies involving patients with acute health conditions tend to conclude that patients can accurately recall past health states (e.g., Singer, Kowalska, & Thode, 2001), whereas studies that included patients with chronic health conditions tend to conclude that they cannot (e.g., Eich, Reeves, Jaeger, & Graff-Radford, 1985).

In her description of "everyday memory," Cohen (2008) lists several other factors that are thought to influence patients' memories of past experiences. Some of these factors – such as personality – have been shown to relate to differences in reporting styles and patients' ability to accurately recall past health states. For example, in a study comparing pre-therapy and post-therapy ratings of pre-intervention symptomatic distress in 65 patients receiving psychotherapy at a university counseling centre (Safer & Keuler, 2002), most patients reported less pre-intervention distress following psychotherapy. The magnitude

of this directional bias was positively correlated with measures of neuroticism and negatively correlated with measures of self-deceptiveness.

2.7 Characteristics of interventions as predictors of discrepancies between ratings

The amount and type of anaesthetic administered during surgery may influence patients' ability to accurately recall past health states. Although few studies have addressed this question directly, other studies have found that patients undergoing general anaesthesia demonstrate larger impairments in cognitive function compared to patients' undergoing local anaesthesia (e.g., Maurer, Chen, Hiebert, Pereira, & Di Cesare, 2007). The duration of anaesthesia and transfusion requirements are also among factors shown to relate to cognitive function following surgery (e.g., Moller, Cluitmans, Rasmussen, Houx, Rasmussen, Canet, Rabbitt, Jolles, Larsen, Hanning, Langeron, Johnson, Lauven, Kristensen, Biedler, van Beem, Fraidakis, Silverstein, Beneken, & Gravenstein, 1998). These impairments in cognitive functioning may include a reduced ability to accurately recall past health states.

Several studies have found patients' receiving different interventions to demonstrate other differences in recall. For example, in a study comparing pre-intervention (prospective) and post-intervention (retrospective) ratings of pre-intervention health and symptom status in patients receiving surgical or non-surgical interventions for impairments of the reproductive system (Aseltine, Carlson, Fowler, & Barry, 2005), measures of change based on pre-intervention (prospective) judgments were significantly

stronger predictors of measures of change based on post-intervention (retrospective) judgments among patients treated non-surgically. In addition, post-intervention health and symptom status were significantly stronger predictors of measures of change based on post-intervention (retrospective) judgments among patients treated surgically. Consistent with Ross's description of implicit theories (1989), the authors suggest that patients treated both surgically and non-surgically may answer a different question than is being posed by the researcher. That is, when asked to compare their pre-intervention health status with their post-intervention health status (after receiving an intervention), both types of patients may instead answer the question: "How healthy are you now?" However, ratings provided by patients treated non-surgically are more likely to agree with true changes in health status. The authors further suggest that the tendency for patients' to exaggerate the benefits of surgery may reflect their cognitive efforts to justify submitting to a stressful treatment.

2.8 Characteristics of trial design as predictors of discrepancies between ratings

The length of the time interval between ratings is also thought to relate to patients' ability to accurately recall past health states. Previous studies have shown the magnitude of the discrepancy between ratings of specific health states made at different points in time to be larger when the time interval is longer. For example, in a study using self-report data from daily pain diaries (McGorry, Webster, Snook, & Hsiang, 1999), differences significant differences in the magnitude of the discrepancy between ratings were found when the time interval between ratings was 6 months, but not when it was 1 week or 1

month. Between-studies comparisons suggest a similar trend to within-studies comparisons. Specifically, the magnitude of the discrepancy between ratings tends to be smaller in studies in which the interval between ratings is 2 weeks or less (e.g., Babul, Darke, Johnson, & Charron-Vincent, 1993; Singer, Kowalska, & Thode, 2001) and larger in studies in which the interval between ratings is greater than 3 months (e.g., Elliot, Smith, Hannaford, Cairns Smith, & Alastair Chambers, 2002; Mancuso & Charlson, 1995).

Previous studies have also demonstrated differences in the patients' ability to accurately recall past health states depending on what constructs were being measured. Specifically, the magnitude of the discrepancy between ratings tends to be larger for general constructs (such as health) than specific constructs (such as pain) (e.g., ten Klooster, Drossaers-Bakker, Taal, & van de Laar, 2007).

2.9 Summary

Chronic musculoskeletal problems are the most common chronic health conditions in Canada. Rigorous evaluation of surgical interventions (through clinical trials using patient-reported outcome measures) is required to decide what constitutes the best care for Canadians with such conditions. Substitution of post-surgical (retrospective) ratings of pre-intervention health status collected for ratings that would have otherwise been collected pre-surgically (prospectively) could increase the efficiency of data collection in clinical trials when patient eligibility cannot be fully determined prior to intra-surgical

evaluation. However, to determine the applicability of this strategy, we require a better understanding of the predictors of discrepancies between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status.

Chapter 3: Objectives

Our primary objective was to examine the predictive relationship between factors that have been shown to relate to patients' ability to accurately recall past states (patient characteristics and aspects of trial design) and discrepancies between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status. In addition to determining whether certain factors were associated with the magnitude of the discrepancy between ratings (the overall degree to which prospective and retrospective ratings differed), we were interested in determining whether these factors were associated with the direction of the discrepancy between ratings (whether, post-surgically, patients tended to rate their pre-intervention health status as being better or worse than reported pre-surgically).

Chapter 4: Methods

4.1 Study design

We combined data from three independent randomized controlled trials conducted in association with researchers from the University of Western Ontario. All three trials were designed to determine patients' ability to accurately recall pre-intervention health status following orthopaedic surgery. Readers can find a detailed description of the methods and results of each trial elsewhere (Bryant, Norman, Stratford, Marx, Walter, & Guyatt, 2006; Gow, 2010; Marsh, Bryant, & MacDonald, 2009). Key design features (assessment schedules and procedures) of these trials are summarized in Table 1.

In each trial, patients were randomized to Group I or Group II. Group I completed assessments before and after surgery. Group II only completed assessments after surgery. At each assessment, patients were asked to rate their current health status on a set of measurement tools. At the first post-surgical assessment, patients were also asked to rate their pre-intervention health status on an identical set of tools.

For the analyses reported in this thesis, we do not include data from Group II and only include data from Group I from the day of surgery (pre-surgical ratings of pre-intervention health status) and the first post-surgical assessment (post-surgical ratings pre-intervention health status and ratings of health status at the time of recall).

Table 1. Assessment schedules and procedures for three independent randomized controlled trials designed to determine patients' ability to accurately recall pre-intervention health status

	Trial I	Trial II	Trial III
Assessment Schedules			
Group I	• four weeks before surgery • day of surgery[a] • two weeks after surgery[b] • one year after surgery	• four weeks before surgery • day of surgery[a] • six weeks after surgery[b] • three months after surgery	• four weeks before surgery • two weeks before surgery • day of surgery[a] • two weeks after surgery[b] • six months after surgery
Group II	• two weeks after surgery • one year after surgery	• six weeks after surgery • three months after surgery	• two weeks after surgery • six months after surgery
Procedures			
Pre-surgical (prospective) ratings of pre-intervention health status	Patients instructed to rate health status according to perceptions of average status during the...		
	• past two weeks	• past four weeks	• past two weeks
Post-surgical (retrospective) ratings of pre-intervention health status	Patients instructed to rate health status according to recollection of average status during the...		
	• two weeks immediately prior to surgery	• four weeks immediately prior to surgery	• two weeks immediately prior to surgery

[a]The pre-surgical (prospective) ratings of pre-intervention health status used in this thesis were collected during this assessment.
[b]The post-surgical (retrospective) ratings of pre-intervention health status used in this thesis were collected during this assessment.

4.2 Measurement of health status

Each trial used different tools to measure health status (see Table 2). Therefore, we combined data from different trials based on the health status construct measured (physical health, mental health, function, or quality of life) instead of the measurement tool used. For the purposes of this thesis, we used the 36-Item Short-Form Health Survey (SF-36) and the 12-Item Short-Form Health Survey (SF-12) to measure physical health and mental health; the Knee Injury and Osteoarthritis Outcome Score (KOOS), Lower Extremity Functional Scale (LEFS), and the Upper Extremity Functional Index (UEFI) to measure function; and the Quality of Life Outcome Measure for Chronic Anterior Cruciate Ligament Deficiency (ACL-QOL), Western Ontario Meniscal Evaluation Tool (WOMET), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Western Ontario Osteoarthritis of the Shoulder Index (WOOS), Western Ontario Rotator Cuff Index (WORC), and Western Ontario Shoulder Instability Index (WOSI) to measure quality of life.

Table 2. Tools used to measure health status in three independent randomized controlled trials designed to determine patients' ability to accurately recall pre-intervention health status

	Measurement Tools
Trial I	• 36-Item Short-Form Health Survey[a] • Knee Injury and Osteoarthritis Outcome Score[b] • International Knee Documentation Committee Subjective Evaluation Form • Quality of Life Outcome Measure for Chronic Anterior Cruciate Ligament Deficiency[c] • Western Ontario Meniscal Evaluation Tool[c]
Trial II	• 12-Item Short-Form Health Survey[a] • Feeling Thermometer • Lower Extremity Functional Scale[b] • Oxford Hip Score • Western Ontario and McMaster Universities Osteoarthritis Index[c]
Trial III	• 36-Item Short-Form Health Survey[a] • Upper Extremity Functional Index[b] • American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form (Patient Report Section) • Western Ontario Osteoarthritis of the Shoulder Index[c] • Western Ontario Rotator Cuff Index[c] • Western Ontario Shoulder Instability Index[c]

[a]Used to measure physical health and mental health.
[b]Used to measure function.
[c]Used to measure quality of life.

The SF-36 (Ware & Sherbourne, 1992) is a 36-item generic measurement tool. Each item has a corresponding ordinal scale with three (10 items), five (25 items), or six (1 item) response options. Scores are reported as a profile of eight domains (Physical Functioning, Role Physical, Bodily Pain, General Health, Vitality, Social Functioning, Role Emotional, and Mental Health) or as two summary measures (PCS and MCS). The SF-36 is used extensively and has demonstrated validity, reliability, and responsiveness in a wide variety of populations, including patients with orthopaedic conditions (McHorney, Ware, & Raczek, 1993; McHorney, Ware, Lu, & Sherbourne, 1994; Shapiro, Richmond, Rockett, McGrath, & Donaldson, 1996; Ware, 2000).

The SF-12 (Ware, Kosinski, & Keller, 1996) is an abbreviated (12-item) version of the SF-36. Scores on this generic measurement tool are reported using the same eight domains and two summary measures as used in the SF-36. The SF-12 is also used extensively and has demonstrated validity, reliability, and responsiveness in a wide variety of populations, including patients with orthopaedic conditions (Luo, George, Kakoura, Edwards, Pietrobon, Richardson, & Hey, 2003; Resnick & Nahm, 2001; Ware, Kosinski, & Keller, 1996; Ware, Kosinski, Turner-Bowker, & Gandek, 2002).

The KOOS (Roos, Roos, Lohmander, Ekdahl, & Beynnon, 1998) is a 42-item site-specific measurement tool for patients with knee problems. Items are grouped into five domains: frequency/severity of physical symptoms (7 items); frequency/severity of pain (9 items); degree of difficulty in activities of daily living (17 items); degree of difficulty in recreational activities (5 items); and quality of life issues (4 items). Each item has a

corresponding ordinal scale with five response options (always, often, sometimes, rarely, never). The KOOS has demonstrated validity, reliability, and responsiveness (Roos & Lohmander, 2003; Roos, Roos, Ekdahl, & Lohmander, 1998; Roos, Roos, Lohmander, Ekdahl, & Beynnon, 1998).

The LEFS (Binkley, Stratford, Lott, & Riddle, 1999) is a 20-item site-specific measurement tool for patients with lower extremity problems. Each item has a corresponding ordinal scale with five response options (0, 1, 2, 3, 4), where zero represents extreme difficulty or inability to perform a specific activity and four represents no difficulty in performing the activity. The LEFS has demonstrated validity, reliability, and responsiveness (Binkley, Stratford, Lott, & Riddle, 1999; Watson, Propps, Ratner, Zeigler, Horton, & Smith, 2005; Yeung, Wessel, Stratford, & MacDermid, 2009).

The UEFI (Stratford, Binkley, & Stratford, 2001) is a 20-item site-specific measurement tool for patients with upper extremity problems. Each item has a corresponding ordinal scale with five response options (0, 1, 2, 3, 4), where zero represents extreme difficulty or inability to perform a specific activity and four represents no difficulty in performing the activity. The UEFI has demonstrated validity, reliability, and responsiveness (Razmjou, Bean, van Osnabrugge, MacDermid, & Holtby, 2006; Stratford, Binkley, & Stratford, 2001).

The ACL-QOL (Mohtadi, 1998) is a 32-item condition-specific measurement tool for patients with anterior cruciate ligament deficiency. Items are grouped into five domains:

physical symptoms (5 items); issues at school/work (4 items); issues with recreation (12 items); lifestyle issues (6 items); and social and emotional issues (5 items). Each item has a corresponding 100 mm visual analogue scale with labeled anchors at 0 mm (e.g., extremely concerned, totally limited, extremely difficult) and 100 mm (e.g., not concern at all, no limitations, not difficult at all). The ACL-QOL has demonstrated validity, reliability, and responsiveness (Mohtadi, 1998).

The WOMET (Kirkley, Griffin, & Whelan, 2007) is a 16-item condition-specific measurement tool for patients with conditions of the meniscus of the knee. Items are grouped into three domains: concern regarding physical symptoms (9 items); issues at school/work or with recreation/lifestyle (4 items); and other issues (3 items). Each item has a corresponding 100 mm visual analogue scale with labeled anchors at 0 mm (e.g., not at all worried) and 100 mm (e.g., extremely worried). The WOMET has demonstrated validity, reliability, and responsiveness (Kirkley, Griffin, & Whelan, 2007).

The WOMAC (Bellamy, 1982) is a 24-item condition-specific measurement tool designed for patients with osteoarthritis. Items are grouped into two domains: severity of physical symptoms (7 items) and degree of difficulty in activities of daily living (17 items). Each item has a corresponding ordinal scale with five response options (none, mild, moderate, severe, extreme). The WOMAC is extensively used and has demonstrated validity, reliability, and responsiveness (Bellamy, Buchanan, Goldsmith, Campbell, & Stitt, 1988; Davies, Watson, & Bellamy, 1999; Roos, Klässbo, & Lohmander, 1999).

The WOOS (Lo, Griffin, & Kirkley, 2001) is a 19-item condition-specific measurement tool designed specifically for patients with osteoarthritis of the shoulder. Items are grouped into four domains: severity of physical symptoms (6 items); concerns regarding function in recreational and work activities (5 items); concerns regarding lifestyle changes (5 items); and emotional issues (3 items). Each item has a corresponding 100 mm visual analogue scale with anchors at 0 mm (e.g., not affected, no difficulty) and 100 mm (e.g., extremely affected, extreme difficulty). The WOOS has demonstrated validity, reliability, and responsiveness (Lo, Griffin, & Kirkley, 2001).

The WORC (Kirkley, Alvarez, & Griffin, 2003) is a 21-item condition-specific measurement tool designed for patients with conditions of the rotator cuff complex of the shoulder. Items are grouped into five domains: severity of physical symptoms (6 items); concerns regarding function in recreational activities (4 items); concerns regarding function in work activities (4 items); concerns regarding lifestyle changes (4 items); and emotional issues (3 items). Each item has a corresponding 100 mm visual analogue scale with anchors at 0 mm (e.g., not affected, no concern, no difficulty) and 100 mm (e.g., extremely affected, extremely concerned, extreme difficulty). The WORC has demonstrated validity, reliability, and responsiveness (Kirkley, Alvarez, & Griffin, 2003; MacDermid, Drosdowech, & Faber, 2006).

The WOSI (Kirkley, Griffin, McLintock, & Ng, 1998) is a 21-item condition-specific measurement tool designed for patients with shoulder instability. Items are grouped into four domains: severity of physical symptoms (10 items); concerns regarding function in

recreational and work activities (4 items); concerns regarding lifestyle changes (4 items); and emotional issues (3 items). Each item has a corresponding 100 mm visual analogue scale with anchors at 0 mm (e.g., no concern, not limited, no difficulty) and 100 mm (e.g., extremely concerned, extremely limited, extreme difficulty). The WOSI has demonstrated validity, reliability, and responsiveness (Kirkley, Griffin, McLintock, & Ng, 1998; Salomonsson, Ahiström, Dalén, & Lillkrona, 2009).

Several types of scores were generated from ratings on these measurement tools for use as independent and dependent variables in our regression analyses. From ratings on the SF-36 and SF-12, we generated one domain score (Mental Health domain score) and two summary scores (the Physical Component Summary (PCS) and the Mental Component Summary (MCS)). From ratings on the KOOS, we generated one domain score (using points allocated to the 17 items related to difficulties in activities of daily living). From ratings on all other measurement tools, we generated summary scores for each tool as described in the literature.

Although we used more than one tool to measure quality of life in Trial I (ACL-QOL and WOMET) and Trial III (WOOS, WORC, and WOSI), each participant only used one of these condition-specific tools. Thus, our final data set included only one pre-surgical (prospective) and one post-surgical (retrospective) rating of pre-intervention physical health, mental health, function, and quality of life (plus one rating of health status at the time of recall for each construct) for each participant. All ratings were transformed to

scales whereby the lowest possible score represented the worst possible health status and the highest possible score represented the best possible health status.

4.3 Eligibility criteria

All patients were scheduled to receive surgical interventions for orthopaedic problems. In Trial I, patients were scheduled to receive arthroscopy for knee problems, such as patellofemoral pain, meniscal injury, and osteoarthritis. In Trial II, patients were scheduled to receive arthroplasty because of osteoarthritis of the hip. In Trial III, patients were scheduled to receive arthroscopy for shoulder problems, such as instability, rotator cuff injury, and osteoarthritis. To confirm that patients in different trials could be considered members of the same population (indicated by a relatively homogenous pattern of scores), we generated scatterplots of the pre-surgical (prospective) ratings of pre-intervention health status versus the post-surgical (retrospective) ratings of pre-intervention health status with points labeled according to source (Trial I, II, or II).

To reduce the impact of learning effects, the original trials excluded patients with previous experience using similar measurement tools. These trials also excluded patients undergoing minor procedures (e.g., manipulation or removal of hardware) and those with rare conditions (e.g., metabolic bone disease) that would usually preclude their involvement in clinical trials, as well as those with no fixed address and those who were unwilling or unable to receive the surgical intervention, participate in all assessments, or use the measurement tools due to concomitant conditions, plans to move outside the

vicinity of the participating centers, major psychiatric illness, cognitive impairment, or inability to speak or understand English.

4.4 Statistical analyses

We used the least squares method of multiple linear regression to construct eight regression models (two for each health status construct) using five independent variables and two dependent variables. The independent variables were age (measured in years), sex, mental state at the time of recall (measured by the Mental Health domain of the SF-36 or SF-12), health status at the time of recall (measured by the same tool used to measure the dependent variable at the time of recall), and time between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status (measured in days). The dependent variables were the magnitude and direction of the discrepancy between pre-surgical (prospective) and post-surgical (retrospective) ratings. The absolute and simple differences between ratings were used to operationalize these variables.

All five independent variables were included in models related to function and quality of life. However, mental state was not included in models related to physical health and mental health, as some of the items used to calculate the Mental Health domain score are the same items used to calculate the MCS and PCS.

Before generating the regression models, we transformed all variables measured by different tools into the same metric by computing standard scores with a mean of 50 and a standard deviation of 10. The difference scores (used as dependent variables) were computed using these standard scores.

We took two precautions to avoid overfitting the regression models and to decrease the probability of spurious findings due to chance. First, instead of applying a more algorithmic and sample-specific method to explore the strength of association between each independent variable and dependent variable (such as stepwise regression), we simultaneously entered all independent variables into each model and retained all independent variables in the models regardless of their significance (direct-entry regression). Second, because we created two different models for each construct, we set a threshold of significance of alpha for each model to < 0.025.

After generating each regression model, we performed graphical and numerical diagnostic tests to check for general problems with our data (collinearity problems and unusual data). To check for collinearity problems, we used tolerance values. To check for unusual data, we used frequency histograms of the standardized residuals, centered leverage values, and two measures of influence (Cook's distance scores and dfbeta values).

We also performed graphical and numerical diagnostic tests (residual analysis) to assess whether our data met the assumptions of linear modeling (linearity of errors,

homoscedasticity of errors, and normality of the error distribution). To assess the assumptions of linearity and homoscedasticity of errors, we used scatterplots of the standardized residuals versus the standardized predicted values. To assess the assumption of the normality of the error distribution, we used Q-Q plots of the standardized residuals, Shapiro-Wilk statistics, and descriptive statistics of the standardized residuals.

Chapter 5: Results

5.1 Patient characteristics

There were 411 patients with complete data for all independent and dependent variables. Table 3 presents the characteristics of patients in the three independent randomized controlled trials. Table 4 presents descriptive statistics for ratings provided on the day of surgery and at the first post-surgical assessment. Table 5 presents descriptive statistics for the absolute and simple differences between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status.

The observed pattern of points on scatterplots of the pre-surgical (prospective) ratings of pre-intervention health status versus the post-surgical (retrospective) ratings of pre-intervention health status confirmed that patients from different trials could be considered members of the same population. Figure 1 illustrates the observed pattern for quality of life. Similar patterns were observed for all other health status constructs.

Table 3. Characteristics of participants in three independent randomized trials controlled trials designed to determine patients' ability to accurately recall pre-intervention health status

	Trial I (N=166)	Trial II (N=115)	Trial III (N=130)	Combined (N=411)
Surgical intervention	knee arthroscopy	hip arthroplasty	shoulder arthroscopy	Trial I (40%) Trial II (28%) Trial III (32%)
Sex (percent male)	63	47	78	63
Age (years)				
Median	40	71	45	49
Range	15 to 78	54 to 90	17 to 83	15 to 90
Height (inches)[a]				
Median	69	66	70	68
Range	60 to 78	56 to 77	60 to 78	56 to 78
Weight (pounds)[b]				
Median	180	181	188	181
Range	115 to 290	69 to 300	120 to 360	69 to 360
Smoking status (percent)[c]				
Never smoked	52	52	54	53
Smoked, but quit	27	34	29	29
Currently smoke	21	14	17	18

[a] Unknown for 18 patients.
[b] Unknown for 18 patients.
[c] Unknown for 3 patients.

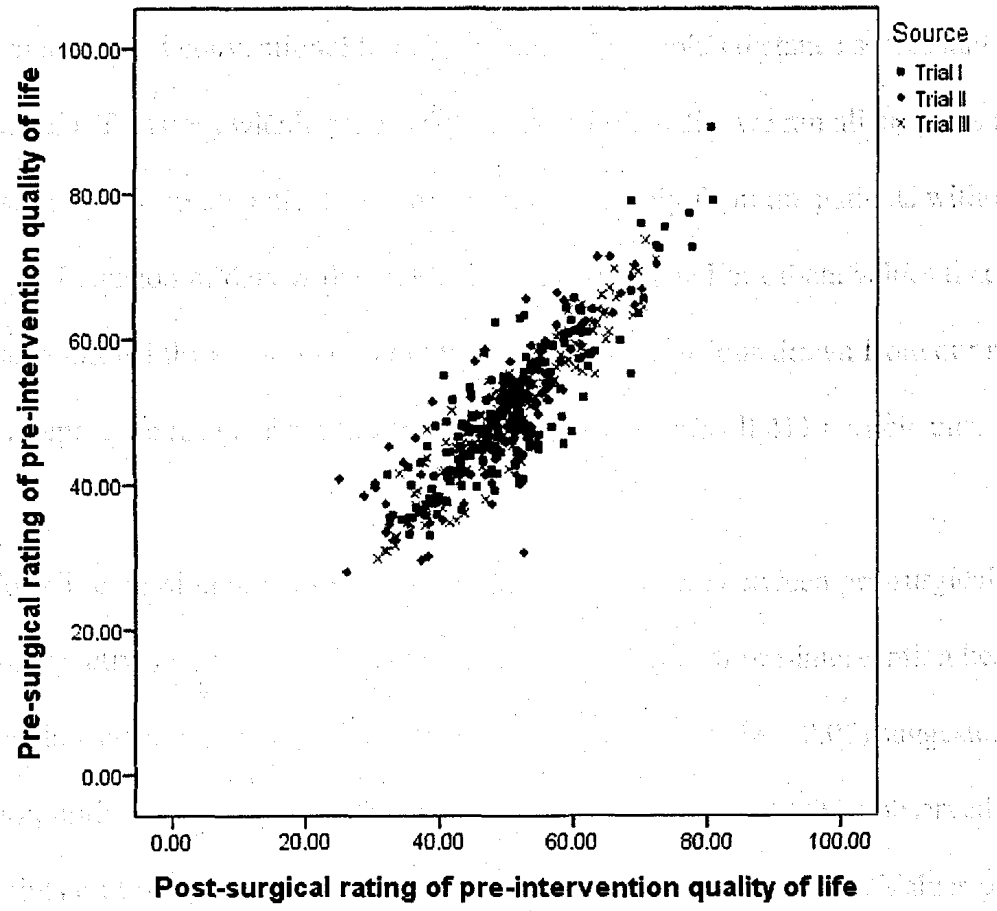
Table 4. Descriptive statistics for ratings provided on the day of surgery and at the first post-surgical assessment

	Construct	Mean	SD	Minimum	Maximum
Day of surgery					
Pre-surgical (prospective) ratings of pre-intervention health status	Physical Health	49.74	9.97	21.69	80.46
	Mental Health	49.96	10.06	20.26	74.29
	Function	49.71	9.96	21.16	76.09
	Quality of Life	49.99	10.05	27.97	89.32
First post-surgical assessment					
Post-surgical (retrospective) ratings of pre-intervention health status	Physical Health	49.95	9.96	25.30	85.91
	Mental Health	49.96	9.96	12.99	71.34
	Function	49.97	9.98	22.40	77.22
	Quality of Life	50.02	9.95	24.84	80.90
Ratings of health status at the time of recall	Physical Health	49.91	9.95	30.18	80.70
	Mental Health	49.95	9.97	19.22	70.36
	Function	49.89	9.96	24.33	80.55
	Quality of Life	49.87	9.81	27.88	84.10

Table 5. Descriptive statistics for the absolute and simple differences between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status

	Absolute Difference (| post-surgical – pre-surgical |)				Simple Difference (post-surgical – pre-surgical)			
	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
Physical Health	5.26	4.87	0.01	28.44	0.21	7.17	- 25.74	28.44
Mental Health	5.89	5.82	0.01	32.24	0.03	8.28	- 32.24	28.27
Function	3.75	4.01	0.00	35.30	0.28	5.48	- 35.30	24.31
Quality of Life	3.72	3.33	0.02	21.82	0.03	5.00	- 15.96	21.82

Figure 1. Scatterplot pre-surgical (prospective) ratings versus post-surgical (retrospective) ratings of pre-intervention quality of life



5.2 Predictors of the magnitude of the discrepancy between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status

Based on measures of influence, we identified participants with unusual data (i.e., data that surpassed conventional threshold values for Cook's distance scores and/or dfbeta values). To assess whether these data were problematic, we ran all analyses twice: once using data from all patients and once using data only from the patients without unusual data. Inclusion of data with Cook's distance scores and/or dfbeta values that surpassed conventional threshold values did not alter the conclusions drawn from our results. Thus, we report the results from the analyses using data from all 411 participants.

For all regression models using the absolute difference between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status as the dependent variable, the results of Shapiro-Wilk tests ($p < 0.05$) suggested that the assumption of normality of the error distribution was not met. The observed pattern of points on Q-Q plots suggested excessive skewness and/or kurtosis. Values of normality statistics suggested that, although the distribution was positively skewed, excessive leptokurtosis was the primary problem with the error distribution. Therefore, after constructing a series of Q-Q plots to estimate the value of p that best corrected for observed departure from normality (for a description of this method, see Tan, Gan, & Chang, 2004), we conducted a power transformation of the dependent variable and re-ran the regression analysis. Follow-up graphical and numerical diagnostic tests using the residuals from the transformed version of the dependent variable verified that the transformed data met the assumptions of linear modeling.

Three of the four models explained a significant amount of variance in the absolute difference between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status (physical health: $R^2 = 0.040$, $p = 0.002$; mental health: $R^2 = 0.114$, $p < 0.001$; function: $R^2 = 0.048$, $p = 0.001$). The model pertaining to quality of life failed to explain a significant amount of variance in the absolute difference between ratings ($R^2 = 0.018$, $p = 0.204$).

Although regression coefficients for each independent variable were statistically significant in one or more models, all partial correlations were small. Table 6 presents the R-square values, regression coefficients, and partial correlations for predictors of the absolute difference between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status.

5.3 Predictors of the direction of the discrepancy between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status

All four models failed to explain a significant amount of variance in the simple difference between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status. Table 7 presents the R-square values, regression coefficients, and partial correlations for predictors of the simple difference between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status.

Table 6. R-square values, regression coefficients, and partial correlations for predictors of the absolute difference between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status

Construct	Predictors	Combined Relationship		Individual Relationship		
		R^2	p	B (95% CI)	p	r
Physical Health	Age	0.040	p=0.002	0.001 (-0.002, 0.004)	p=0.527	0.031
	Sex			0.095 (0.001, 0.189)	p=0.047	0.098
	Physical Health at Time of Recall			0.001 (-0.003, 0.006)	p=0.617	0.025
	Time Between Ratings			0.004 (0.001, 0.007)	p=0.023	0.113
Mental Health	Age	0.114	p<0.001	0.005 (0.002, 0.008)	p=0.001	0.170
	Sex			0.188 (0.090, 0.287)	p<0.001	0.184
	Mental Health at Time of Recall			-0.008 (-0.012, -0.003)	p=0.001	-0.160
	Time Between Ratings			0.000 (-0.003, 0.004)	p=0.942	0.004
Function	Age	0.048	p=0.001	0.004 (0.001, 0.007)	p=0.006	0.136
	Sex			0.018 (-0.074, 0.110)	p=0.707	0.019
	Mental State at Time of Recall			0.000 (-0.004, 0.005)	p=0.948	0.003
	Function at Time of Recall			0.002 (-0.002, 0.007)	p=0.295	0.052
	Time Between Ratings			0.002 (-0.002, 0.005)	p=0.304	0.051
Quality of Life	Age	0.018	p=0.204	0.003 (0.000, 0.005)	p=0.049	0.097
	Sex			-0.006 (-0.091, 0.079)	p=0.884	-0.007
	Mental State at Time of Recall			0.002 (-0.002, 0.006)	p=0.348	0.047
	Quality of Life at Time of Recall			-0.001 (-0.005, 0.003)	p=0.747	-0.016
	Time Between Ratings			0.000 (-0.003, 0.003)	p=0.884	0.007

Table 7. R-square values, regression coefficients, and partial correlations for predictors of the simple difference between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status

Construct	Predictors	Combined Relationship		Individual Relationship		
		R^2	p	B (95% CI)	p	r
Physical Health	Age	0.019	p=0.091	-0.036 (-0.080, 0.009)	p=0.116	-0.078
	Sex			1.424 (-0.051, 2.899)	p=0.058	0.094
	Physical Health at Time of Recall			0.058 (-0.012, 0.129)	p=0.105	0.080
	Time Between Ratings			0.002 (-0.050, 0.054)	p 0.945	0.003
Mental Health	Age	0.022	p 0.057	-0.032 (-0.083, 0.020)	p 0.224	-0.060
	Sex			0.422 (-1.271, 2.116)	p 0.624	0.024
	Mental Health at Time of Recall			0.114 (0.033, 0.194)	p=0.006	0.137
	Time Between Ratings			0.030 (-0.029, 0.090)	p=0.315	0.050
Function	Age	0.027	p=0.050	-0.039 (-0.073, -0.005)	p=0.026	-0.110
	Sex			1.297 (0.167, 2.426)	p=0.025	0.111
	Mental State at Time of Recall			0.031 (-0.024, 0.087)	p 0.271	0.055
	Function at Time of Recall			-0.012 (-0.069, 0.045)	p 0.681	-0.020
	Time Between Ratings			0.003 (-0.037, 0.043)	p=0.867	0.008
Quality of Life	Age	0.008	p=0.662	0.007 (-0.024, 0.039)	p=0.650	0.023
	Sex			0.250 (-0.786, 1.287)	p=0.635	0.024
	Mental State at Time of Recall			-0.039 (-0.088, 0.011)	p=0.126	-0.076
	Quality of Life at Time of Recall			0.022 (-0.029, 0.073)	p 0.395	0.042
	Time Between Ratings			-0.002 (-0.038, 0.034)	p 0.911	-0.006

Chapter 6: Discussion

The results of this study suggest that, although age, sex, mental state at the time of recall, health status at the time of recall, and time between ratings are related to discrepancies between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status, they are not strong predictors of the magnitude and direction of this discrepancy in the early post-surgical period in patients receiving surgical interventions for common orthopaedic problems. This was demonstrated when independent variables were considered in combination and individually. It was also consistent across a variety of health status constructs, including physical health, mental health, function, and quality of life.

In multiple regression, the R-square value represents how well a model fits the data by indicating the proportion of variance in the dependent variable explained by all independent variables within the model. Regression coefficients represent the unique contribution of each independent variable to the prediction of the dependent variable. Partial correlations represent the relationship between each independent variable and a dependent variable after controlling for all other independent variables in the model. Although regression coefficients and partial correlations are mathematically related to each other and yield similar information, the latter are more easily interpreted because they are scaleless and, when squared, specify the unique proportion of variance in the dependent variable explained by a particular independent variable.

Although several R-square values and regression coefficients were statistically significant, the small size of the R-square values and partial correlations suggested that – in combination and individually – the independent variables examined in this study did not demonstrate the "substantive significance" required to be considered strong predictors of discrepancies between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status. In the three models with statistically significant R-square values, age, sex, mental state at the time of recall, health status at the time of recall, and time between ratings – in combination – only explained 4%, 11%, and 5% of the variance in the absolute difference between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention physical health, mental health, and function. Furthermore, individually, predictors explained a maximum of 3% of the variance in the absolute difference between ratings. This means that the amount of variance in the dependent variable unaccounted for by the independent variables in our models is, at minimum, 89% (when independent variables are considered in combination) or 97% (when independent variables are considered individually).

The total variance in any variable can be divided into two types: systematic (which is attributable to specific factors) and unsystematic (which is random). In multiple regression, the ratio of systematic to unsystematic error increases with each independent variable added to a model (provided the strength of the relationship between the independent variable and dependent variable is measurable given the nature of the variables examined and measurement tools used). The R-square value also increases. The small R-square values in this study, thus, suggest that there are other factors – some

identifiable, some not – that account for the remainder of the variance in the magnitude and direction of the discrepancy between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status unaccounted for by age, sex, mental state at the time of recall, health status at the time of recall, and time between ratings.

In her description of "everyday memory," Cohen (2008) encourages researchers to recognize that people's memories for past experiences are influenced by a multitude of factors, including other past experiences, culture, current motives, personality, and future plans. Previous studies suggest that these and other factors – such as the nature of a patient's health condition (chronic versus acute) and the nature of the intervention – are also related to discrepancies between ratings of specific health states made at different points in time. A limitation of this study is that we did not include these factors in our regression models. While it is probably impossible (and definitely impractical) to take all of these factors into account, it is possible that one or more of these factors are a strong predictor of discrepancies between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status in patients receiving surgical interventions for common orthopaedic conditions.

Interestingly, although not "substantively significant," the statistically significant trends observed in this study are consistent with those shown in other studies. That is, other studies have shown the magnitude of the discrepancy between specific health states made at different points in time to be larger when the time between ratings is larger (e.g.,

McGorry, Webster, Snook, & Hsiang, 1999) and to be larger in those who are in a worse health state at the time of recall (e.g., Jamison, Sbrocco, & Parris, 1989), are older (e.g., Lingard, Wright, Sledge, & the Kinemax Outcomes Group, 2001), or are female (e.g., Hunter, Philips, & Rachman, 1979). Other studies have also shown patients who are in a better health state at the time of recall (e.g., Kent, 1985) or are female (e.g., Mancuso & Charlson, 2005) to tend to rate their pre-intervention health status as being better after receiving the intervention.

As noted by Ross (1989), the majority of studies that have reported a significant difference between ratings of specific health states made at different points in time were designed to emphasize the processes by which people construct memories and, if the focus shifts to comparing ratings, the differences are relatively small. This is also consistent with the mean differences between ratings observed in our study, which ranged from 3.72 to 5.89 (absolute difference) or 0.03 to 0.28 (simple difference).

Strengths of this study include its sound design, large sample, and rigorous statistical plan. The potential argument that the observed level of agreement between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status was overly influenced by repeated exposure to measurement tools was addressed by the design of the three randomized controlled trials from which data for the present study were derived (Bryant, Norman, Stratford, Marx, Walter, & Guyatt, 2006; Gow, 2010; Marsh, Bryant, & MacDonald, 2009). In each of these studies, there were no important between-groups differences in post-surgical (retrospective) ratings of pre-intervention

health status. By combining data sets from these studies, we obtained data for a total of 411 patients receiving a variety of surgical interventions for common knee, hip, and shoulder problems. This enhances the applicability of our findings and addresses the potential criticism that we included too many independent variables in our models (as, to avoid the problem of overfitting, a ratio of at least ten observations per independent variable is generally recommended) (Babyak, 2004). To decrease the probability of spurious findings due to chance, we selected all independent variables *a priori* based on a thorough literature review that identified characteristics of patients and trial design previously shown to relate to discrepancies between ratings of specific health states made at different points in time, simultaneously entered all independent variables into each model, and retained all independent variables in the models regardless of their significance, and adjusted our threshold of significance of alpha for R-square values to reflect the fact that we created two regression models for each construct. These precautions enhance our confidence that similar results would be found if this study were repeated in a similar population.

Chapter 7: Conclusion

Individually and in combination, age, sex, mental state at the time of recall, health status at the time of recall, and time between ratings are not strong predictors of the magnitude or direction of the discrepancy between pre-surgical (prospective) and post-surgical (retrospective) ratings of pre-intervention health status in the early post-surgical period in patients receiving surgical interventions for common orthopaedic problems. Future research should examine the predictive relationship between other factors and discrepancies between pre-surgical and post-surgical ratings of pre-intervention physical health, mental health, function, and quality of life in this population.

References

Advisory Committee on Population Health. (2002). *Advancing integrated prevention strategies in Canada: An approach to reducing the burden of chronic diseases.* Ottawa: Queen's Printer.

Aseltine, R.H., Jr., Carlson, K.J., Fowler, F.J., Jr., & Barry, M.J. (1995). Comparing prospective and retrospective measures of treatment outcomes. *Medical Care, 33*(4), AS67-AS76, Supplement.

Babul, N., Darke, A.C., Johnson, D.H., & Charron-Vincent, K. (1993). Using memory for pain in analgesic research. *Annals of Pharmacotherapy, 27*(1), 9-12.

Babyak, M.A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411-421.

Beaton, D.E., & Schemitsch, E. (2003). Measures of health-related quality of life and physical function. *Clinical Orthopaedics and Related Research, 413*, 90-105.

Bellamy, N. (1982). *Osteoarthritis – an evaluative index for clinical trials* (Master's Thesis). McMaster University, Hamilton, Ontario, Canada.

Bellamy, N., Buchanan, W.W., Goldsmith, C.H., Campbell, J., & Stitt, L.W. (1988). Validation study of WOMAC: A health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoporosis of the hip or knee. *Journal of Rheumatology, 15*(12), 1833-1840.

Binkley, J.M., Stratford, P.W., Lott, S.A., & Riddle, D.L., for the The North American Orthopaedic Rehabilitation Research Network. (1999). The Lower Extremity Functional Scale (LEFS): Scale development, measurement properties, and clinical application. *Physical Therapy*, 79(4), 371-383.

Borrell-Carrió, F., Suchman, A.L., & Epstein, R.M. (2004). The biopsychosocial model 25 years later: Principles, practice, and scientific inquiry. *Annals of Family Medicine, 2*(6), 576-582.

Bryant, D. (2005). *Improving the efficiency in collection of pre-operative self-ratings of quality of life, functional status and general health in patients undergoing knee surgery* (Doctoral Thesis). McMaster University, Hamilton, Ontario, Canada.

Bryant, D., Dill, J., Litchfield, R., Amendola, A., Giffin, R., Fowler, P., & Kirkley, A. (2007). Effectiveness of bioabsorbable arrows compared with inside-out suturing

for vertical, reparable meniscal lesions: A randomized clinical trial. *American Journal of Sports Medicine*, *35*(6), 889-896.

Bryant, D., & Fernandes, N. (2011). Measuring patient outcomes: A primer. *Injury*, *42*(3), 232-235.

Bryant, D., Litchfield, R., Holtby, R., Willits, K., Drosdowech, D., Spouge, A., & Guyatt, G. (n.d.). A randomized clinical trial to compare the effectiveness of rotator cuff repair with or without augmentation using porcine small intestine submucosa (SIS) for patients with moderate to large rotator cuff tears: A pilot study. Manuscript in preparation.

Bryant, D., Norman, G., Stratford, P., Marx, R.G., Walter, S.D., & Guyatt G. (2006). Patients undergoing knee surgery provided accurate ratings of preoperative quality of life and function 2 weeks after surgery. *Journal of Clinical Epidemiology*, *59*(9), 984-993.

Bryant, D., Schünemann, H., Brozek, J., Jaeschke, R., & Guyatt, G. (2007). Patient reported outcomes: General principles of development and interpretability. *Polish Archives of Internal Medicine*, *117*(4), 5-11.

Busse, J.W., & Heetveld, M.J. (2006) Critical appraisal of the orthopaedic literature: Therapeutic and economic analysis. *Injury*, *37*(4), 312-320.

Canadian Institutes of Health Research. (n.d.) *Glossary of funding related terms*. Retrieved May 30, 2011, from http://www.cihr-irsc.gc.ca/e/34190.html.

Canadian Orthopaedic Care Strategy Group. (2010). *A backgrounder report from the Canadian Orthopaedic Care Strategy Group: Building a collective policy agenda for musculoskeletal health and mobility*. Toronto: Arthritis Community Research and Evaluation Unit.

Chung, K.C., & Burns, P.B. (2008). Guide to planning and executing a surgical randomized controlled trial. *Journal of Hand Surgery*, *33-A*(3), 407-412.

Cohen, G. (2008). The study of everyday memory. In G. Cohen & M.A. Conway (Eds.), *Memory in the real world* (Third Edition). New York: Psychology Press.

Davies, G.M., Watson, D.J., & Bellamy, N. (1999). Comparison of the responsiveness and relative effect size of the Western Ontario and McMaster Universities Osteoarthritis Index and the Short-Form Medical Outcomes Study Survey in a randomized, clinical trial of osteoarthritis patients. *Arthritis Care and Research*, *12*(3), 172-179.

Dawson, E.G., Kanim, L.E., Sra, P., Dorey, F.J., Goldstein, T.B., Delamarter, R.B., & Sandhu, H.S. (2002). Low back pain recollection versus concurrent accounts: Outcomes analysis. *Spine*, *27*(9), 984-993.

Eich, E., Reeves, J.L., Jaeger, B., & Graff-Radford, S.B. (1985). Memory for pain: Relation between past and present pain intensity. *Pain*, *23*(4), 375-380.

Elliot, A.M., Smith, B.H., Hannaford, P.C., Cairns Smith, W., & Alastair Chambers, W. (2002). Assessing change in chronic pain severity: The chronic pain grade compared with retrospective perceptions. *British Journal of General Practice*, *52*(477), 269-274.

Everts, B., Karlson, B., Währborg, P., Abdon, N., Herlitz, J., & Hedner, T. (1999). Pain recollection after chest pain of cardiac origin. *Cardiology*, *92*(2), 115-120.

Gow, J. (2010). *Methodological efficiency in orthopedic trials* (Doctoral Thesis). The University of Western Ontario, London, Ontario, Canada.

Hunter, M., Philips, C., & Rachman, S. (1979). Memory for pain. *Pain*, *6*(1), 35-46.

Jamison, R.N., Sbrocco, T., & Parris, W.C.V. (1989). The influence of physical and psychosocial factors on accuracy of memory for pain in chronic pain patients. *Pain*, *37*(3), 289-294.

Kent, G. (1985). Memory of dental pain. *Pain*, *21*(2), 187-194.

Kirkley, A., Alvarez, C., & Griffin, S. (2003). The development and evaluation of a disease-specific quality-of-life questionnaire for disorders of the rotator cuff: The Western Ontario Rotator Cuff Index. *Clinical Journal of Sport Medicine*, *13*(2), 84-92.

Kirkley, A., Griffin, S., McLintock, H., & Ng, L. (1998). The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability: The Western Ontario Shoulder Instability Index (WOSI). *American Journal of Sports Medicine*, *26*(6), 764-772.

Kirkley, A., Griffin, S., & Whelan, D. (2007). The development and validation of a quality of life measurement tool for patients with meniscal pathology: The Western Ontario Meniscal Evaluation Tool (WOMET). *Clinical Journal of Sport Medicine*, *17*(5), 349-356.

Lingard, E.A., Wright, E.A., Sledge, C.B., & the Kinemax Outcomes Group. (2001). Pitfalls of using patient recall to derive preoperative status in outcomes of total knee arthroplasty. *Journal of Bone and Joint Surgery*, *83-A*(8), 1149-1156.

Linton, S.J. (1991). Memory for chronic pain intensity: Correlates of accuracy. *Perceptual and Motor Skills*, *72*(3), 1091-1095.

Lo, I.K.Y., Griffin, S., & Kirkley, A. (2001). The development of a disease-specific quality of life measurement tool for osteoarthritis of the shoulder: The Western Ontario Osteoarthritis of the Shoulder (WOOS) Index. *Osteoarthritis and Cartilage*, *9*(8), 771-778.

Luo, X., George, M.L., Kakouras, I., Edwards, C.L., Pietrobon, R., Richardson, W., & Hey, L. (2003). Reliability, validity, and responsiveness of the short form 12-item survey (SF-12) in patients with back pain. *Spine*, *28*(15), 1739-1745.

MacDermid, J.C., Drosdowech, D., & Faber, K. (2006). Responsiveness of self-report scales in patients recovering from rotator cuff surgery. *Journal of Shoulder and Elbow Surgery*, *15*(4), 407-414.

MacDermid, J.C., Holtby, R., Razmjou, H., Bryant, D., & JOINTS Canada. (2006). All-arthroscopic versus mini-open repair of small or moderate-sized rotator cuff repairs: A protocol for a randomized trial [NCT00128076]. *BMC Musculoskeletal Disorders*, *7*, 25.

Mancuso, C.A., & Charlson, M.E. (1995). Does recollection error threaten the validity of cross-sectional studies of effectiveness? *Medical Care*, *33*(4), AS77-AS88.

Marsh, J., Bryant, D., & MacDonald, S. (n.d.) Age, gender, mental health, time between ratings, anaesthetic type, duration of anaesthesia, and current health status are not significant predictors of ability to recall preoperative quality of life, function, and general health status following total hip arthroplasty. Submitted to the *Journal of Bone and Joint Surgery*.

Marsh, J., Bryant, D., & MacDonald, S.J. (2009). Older patients can accurately recall their preoperative health status six weeks following total hip arthroplasty. *Journal of Bone and Joint Surgery*, *91-A*(12), 2827-2837.

Maurer, S.G., Chen, A.L., Hiebert, R., Pereira, G.C., & Di Cesare, P.E. (2007). Comparison of outcomes of using spinal versus general anesthesia in total hip arthroplasty. *American Journal of Orthopedics*, *36*(7), E101-E106.

McGorry, R.W., Webster, B.S., Snook, S.H., & Hsiang, S.M. (1999). Accuracy of pain recall in chronic and recurrent low back pain. *Journal of Occupational Rehabilitation*, *9*(3), 169-178.

McHorney, C.A., Ware, J.E., Jr., & Raczek, A.E. (1993). The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care*, *31*(3), 247-263.

McHorney, C.A., Ware, J.E., Jr., Lu, J.F.R., & Sherbourne, C.D. (1994). The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care, 32*(1), 40-66.

Mohtadi, N. (1998). Development and validation of the quality of life outcome measure (questionnaire) for chronic anterior cruciate ligament deficiency. *American Journal of Sports Medicine, 26*(3), 350-359.

Moller, J.T., Cluitmans, P., Rasmussen, L.S., Houx, P., Rasmussen, H., Canet, J., Rabbitt, P., Jolles, J., Larsen, K., Hanning, C.D., Langeron, O., Johnson, T., Lauven, P.M., Kristensen, P.A., Biedler, A., van Beem, H., Fraidakis, O., Silverstein, J.H., Beneken, J.E.W., & Gravenstein, J.S., for the ISPOCD investigators. (1998). Long-term postoperative cognitive dysfunction in the elderly: ISPOCD1 study. *Lancet, 351*(9106), 857-861.

Murphy, K.A., Spence, S.T., McIntosh, C.N., & Connor Gorber, S.K., for the Population Health Impact of Disease in Canada (PHI). (2006). *Health state descriptions for Canadians: Musculoskeletal diseases.* Ottawa: Statistics Canada.

Nasmith, L., Ballem, P., Baxter, R., Bergman, H., Colin-Thomé, D., Herbert, C., Keating, N., Lessard, R., Lyons, R., McMurchy, D., Ratner, P., Rosenbaum, P., Tamblyn, R., Wagner, E., & Zimmerman, B. (2010). *Transforming care for Canadians with chronic health conditions: Put people first, expect the best, manage for results.* Ottawa: Canadian Academy of Health Sciences.

Portney, L.G., & Watkins, M.P. (2009). *Foundation of clinical research: Applications to practice* (Third Edition). Upper Saddle River: Pearson Education.

Razmjou, H., Bean, A., van Osnabrugge, V., MacDermid, J.C., & Holtby, R. (2006). Cross-sectional and longitudinal construct validity of two rotator cuff disease-specific outcome measures. *BMC Musculoskeletal Disorders, 7*, 26.

Resnick, B., & Nahm, E.S. (2001). Reliability and validity testing of the revised 12-Item Short-Form Health Survey in older adults. *Journal of Nursing Measurement, 9*(2), 151-161.

Roos, E.M., Klässbo, M., & Lohmander, L.S. (1999). WOMAC Osteoarthritis Index – reliability, validity, and responsiveness in patients with arthroscopically assessed osteoarthritis. *Scandinavian Journal of Rheumatology, 28*(4), 210-215.

Roos, E.M., & Lohmander, L.S. (2003). The Knee Injury and Osteoarthritis Outcome Score (KOOS): From joint injury to osteoarthritis. *Health and Quality of Life Outcomes, 1*, 64.

Roos, E.M., Roos, H.P., Ekdahl, C., & Lohmander, L.S. (1998). Knee injury and osteoarthritis outcome score (KOOS) – validation of a Swedish version. *Scandinavian Journal of Medicine and Science in Sports, 8*(6), 439-448.

Roos, E.M., Roos, H.P., Lohmander, L.S., Ekdahl, C., & Beynnon, B.D. (1998). Knee Injury and Osteoarthritis Outcome Score (KOOS) – development of a self-administered outcome measure. *Journal of Orthopaedic and Sports Physical Therapy, 78*(8), 88-96.

Ross, M. (1989). Relation of implicit theories to the construction of personal histories. *Psychological Review, 96*(2), 341-357.

Sackett, D.L., Rosenberg, W.M.C., Muir Gray, J.A., Haynes, R.B., & Richardson, W.S. (1996). Evidence based medicine: What it is and what it isn't. *British Medical Journal, 312*(7023), 71-72.

Safer, M.A., & Keuler, D.J. (2002). Individual differences in misremembering pre-psychotherapy distress: Personality and memory distortion. *Emotion, 2*(2), 162-178.

Salomonsson, B., Ahiström, S., Dalén, N., & Lillkrona, U. (2009). The Western Ontario Shoulder Instability Index (WOSI): Validity, reliability, and responsiveness retested with a Swedish translation. *Acta Orthopaedica, 80*(2), 233-238.

Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research, 11*(3), 193-205.

Shapiro, E.T., Richmond, J.C., Rockett, S.E., McGrath, M.M., & Donaldson, W.R. (1996). The use of a generic, patient-based health assessment (SF-36) for evaluation of patients with anterior cruciate ligament injuries. *American Journal of Sports Medicine, 24*(2), 196-200.

Singer, A.J., Kowalska, A., & Thode, H.C., Jr., (2001). Ability of patients to accurately recall the severity of acute painful events. *Academic Emergency Medicine, 8*(3), 292-295.

Sprangers, M.A.G., & Aaronson, N.K. (1992). The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: A review. *Journal of Clinical Epidemiology, 45*(7), 743-760.

Sprangers, M.A.G., & Schwartz C.E. (1999). Integrating response shift into health-related quality of life research: A theoretical model. *Social Science and Medicine, 48*(11), 1507-1515.

Stratford, P.W., Binkley, J.M., & Stratford, D.M. (2001). Development and initial validation of the Upper Extremity Functional Index. *Physiotherapy Canada, 53*(4), 259-267.

Tan, W.D., Gan, F.F., & Chang, T.C. (2004). Using normal quantile plot to select an appropriate transformation to achieve normality. *Computational Statistics and Data Analysis, 45*(3), 609-619.

ten Klooster, P.M., Drossaers-Bakker, K.W., Taal, E., & van de Laar, M.A.F.J. (2007). Can we assess baseline pain and global health retrospectively? *Clinical and Experimental Rheumatology, 25*(2), 176-181.

Ware, J.E., Jr. (2000). SF-36 Health Survey update. *Spine, 25*(24), 3130-3139.

Ware, J.E., Jr., & Sherbourne, C.D. (1992). The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Medical Care, 30*(6), 473-483.

Ware, J.E., Jr., Kosinski, M., & Keller, S.D. (1996). A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*(3), 220-233.

Ware, J.E., Jr., Kosinski, M., Turner-Bowker, D.M., & Gandek, B. (2002). *User's manual for the SF-12v2™ Health Survey*. Lincoln: QualityMetric Incorporated.

Watson, C.J., Propps, M., Ratner, J., Zeigler, D.L., Horton, P., & Smith, S.S. (2005). Reliability and responsiveness of the lower extremity functional scale and the anterior knee pain scale in patients with anterior knee pain. *Journal of Orthopaedic and Sports Physical Therapy, 35*(3), 136-146.

World Health Organization. (2001). *ICF: International Classification of Functioning, Disability and Health*. Geneva: World Health Organization.

World Health Organization. (2005). *Preventing chronic diseases: A vital investment*. Geneva: World Health Organization.

Yeung, T.S., Wessel, J., Stratford, P., & MacDermid, J. Reliability, validity, and responsiveness of the Lower Extremity Functional Scale for inpatients of an orthopaedic rehabilitation ward. *Journal of Orthopaedic and Sports Physical Therapy, 39*(6), 468-477.

Zonneveld, L.N., McGrath, P.J., Reid, G.J., & Sorbi, M.J. (1997). Accuracy of children's pain memories. *Pain, 71*(3), 297-302.